Terminology finding in the Sketch Engine: an Evaluation

Adam Kilgarriff
Lexical Computing Ltd.

ABSTRACT

The Sketch Engine is a leading corpus query tool, in use for lexicography at OUP, CUP, Collins and Le Robert, and at national language institutes of eight countries, and for teaching and research in many universities. Its distinctive feature is the 'word sketch' a one page, automatic, corpus, derived summary of a word's grammatical and collocational behaviour.

Very large corpora and word sketches are available for sixty languages.

A number of tools and resources have recently been added with translators and terminologists in mind. The resources are parallel corpora: EUROPARL-7 and the various datasets available in the OPUS collection. The tools are bilingual word sketches and the term finder. These have been reported on in previous Asling/Aslib conferences.

One remaining task is to make the Sketch Engine functions conveniently accessible to translators and terminologists. We have recently done this via IntelliWebSearch, a tool which lets the user highlight text in the environment they are working in, which could be a CAT tool or Microsoft Word, and, with a key sequence, query their preferred database. So now the key sequence can take the translator or terminologist to a browser window showing the word sketch, or parallel concordance, or any of a number of other reports, for the expression they are working on.

1. The term-finding functionality

The term-finder starts from a domain corpus, and a reference corpus. First it finds all the noun phrases, and their frequencies, in both corpora. It then takes the ratio, and the items with highest ratios will be terms, as in Figure 1 (data supplied by lead users, the World Intellectual Property Organisation).

Term	Frequency	Freq/mill	Score
station de base	28612	3292.2	3293.2
station mobile	12514	1439.9	1440.9
communication sans fil	8189	942.3	943.3
liaison montante	6561	754.9	737.5
terminal mobile	7406	852.2	709.8
liaison descendante	5434	625.3	626.3
stations de base	5010	576.5	577.5
réseau de communication	4255	489.6	490.6
communication mobile	4722	543.3	462.5
point d' accès	3907	449.6	450.6
modes de réalisation	3486	401.1	402.1
réseau d' accès	3241	372.9	373.9
réseau sans fil	2903	334.0	335.0
accès radio	2412	277.5	278.5
transfert intercellulaire	2408	277.1	278.1

14

Figure 1: French terms in the mobile communications domain

In some cases, as with WIPO, the user will have domain corpora, but in others they will not. In that case they may use the BootCaT procedure (Baroni and Bernardini 2004). The user, typically a translator working in a domain where they are not an expert, inputs a few domain-specific 'seed words'; these are sent to a search engine, and the hits identified by the search engine are gathered, cleaned, de-duplicated and processed to give a domain-specific corpus. This functionality has been found to support translators well (Bernardini et al 2013). For some time, the Sketch Engine has incorporated a BootCaT tool, allowing users to create an instant corpus for a domain, which means they can then compare this corpus with a reference corpus to find the keywords of the domain. The functionality has recently been extended so the user can find the terms alongside key words. Thus, where the user has Bootcatted an English environment corpus, the Sketch Engine provides the "key words and terms" report shown in Figure 2.

The requirements for the term-finding functionality are:

- a processing chain, comprising tokeniser, lemmatiser and part-of-speech tagger, installed and ready to apply to the user's domain corpus
- a reference corpus processed with the processing chain
- a term grammar.

At time of writing, these are in place for Chinese, English, French, German, Japanese, Korean, Russian, Spanish and Portuguese.

Figure 2: English keywords and terms in the environment domain. The tickboxes are so the user can iterate the procedure to extend and refine the corpus

2. Evaluation

To evaluate a term-finder for a language and a domain, a list of all the 'true terms' is required. Then we can compute precision and recall.

One problem: how to define the domain? The straightforward answer: provide a corpus of it. Then we have the more constrained task of assessing recall and precision, from a given corpus, when the terms in that corpus are known.

Another problem: won't two different terminologists inevitably deliver two different lists?

We approached the task by hunting for research datasets comprising domain corpora and term-lists derived, by human experts, from them. In most cases, this had been done as part of a term-finding task, so there were also published papers, with term-finding results, over these datasets, so we had a reference result to compare our results with. This addressed the second problem, as, whatever the lists, we were confronted with the same challenge as the resource developers. We found datasets for seven languages and six domains. In each case we entered the corpus into the Sketch Engine, ran the term-finder, and computed precision and recall (which we could then compare with the performance figures of the group who developed the dataset.) The paper will present these results.

3. In sum

The Sketch Engine has for some years been a leading tool for lexicography and corpus linguistics. Its terminology function is now a year old. We present a thorough evaluation.

References

M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004, Lisbon: ELDA. 1313-1316.

S. Bernardini, A. Ferraresi and E. Zanchetta. 2013. Old needs, new solutions: comparable corpora for language professionals. In Sharoff, S., R. Rapp, P. Zweigenbaum, P. Fung, editors. Building and Using Comparable Corpora. Springer